IT550: Information Retrieval

Assignment 4: Stemming

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn about hierarchical clustering, distance measures for measuring the closeness of words and also implement a stemmer from scratch.

1 Problem description

In linguistic morphology and information retrieval, stemming is the process of reducing inflected words to their word stem, base or root form—generally a written word form. Stemming algorithms are commonly called stemmers. A stemming algorithm reduces words 'likely', 'likes', 'liked', 'liking' to its root form 'like'.

2 Implementation

2.1 Dataset

- For this assignment you are provided a file which contains list of words.
- Download the words using the link: https://drive.google.com/open?id=1YOjtv8iyXOPISHmLyAabhl1IibnLcCx_
- For retrieval task use the dataset used in the previous assignment

2.2 Exercise

1. You will first implement YASS stemmer.
2. For given list of word compute their string distance using D1, D2, D3, D4 and Levenstein distance. (Information about the YASS stemmer can be found in the paper)
3. For computed distance perform hierarchical clustering (use any of the measurement single linkage, average linkage and complete linkage) and find center word which would be stem word for all other words in cluster.
4. Generate a graph of no. of clusters vs threshold at which cluster merging is stopped.
5. Use Terrier to index the files and then perform the retrieval task
6. Compare difference in MAP using YASS stemmer and Porter stemmer.

3 References

- YASS yet another suffix stemmer
- https://dzone.com/articles/the-levenshtein-algorithm-1
- https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
- http://terrier.org/
- https://github.com/terrier-org/terrier-core

4 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- Your notebook should contain the graph of no. of clusters vs threshold
- Show the MAP scores for both Porter and YASS stemmer
- The submission deadline for this assignment in **23rd Feb 2020 at 11 PM**